Full-Length Transcriptome Sequences by A Combination of Sequencing Platforms Applied to Isoflavonoid and Triterpenoid Saponin Biosynthesis of Astragalus Mongholicus Bunge

Minzhen Yin
Anhui University of Chinese Medicine

Shanshan Chu
Anhui University of Chinese Medicine

Tingyu Shan
Anhui University of Chinese Medicine

Liangping Zha
Anhui University of Chinese Medicine

Huasheng Peng (✉ hspeng@126.com)
Anhui University of Chinese Medicine  https://orcid.org/0000-0001-7539-2377

Research

Keywords: Astragalus mongholicus Bunge, Isoflavonoids, Triterpenoid saponins, Transcriptome, Biosynthesis

DOI: https://doi.org/10.21203/rs.3.rs-127368/v1

License: ☑️ 📧 This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

**Background:** *Astragalus mongholicus* Bunge is an important medicinal plant and has been used in traditional Chinese medicine for a long history, which is rich in isoflavonoids and triterpenoid saponins. Although these active constituents in *A. mongholicus* have been discovered for a long time, the molecular genetic basis of the isoflavonoid and triterpenoid saponin biosynthesis pathways is virtually unknown due to the lack of a reference genome. The combination of next-generation sequencing (NGS) and single-molecule real-time (SMRT) sequencing to analyze genes involved in the biosynthetic pathways of secondary metabolites in medicinal plants has been widely recognized.

**Results:** In this study, NGS, SMRT sequencing, and targeted compounds were combined to investigate the association between isoflavonoids and triterpenoid saponins and gene expression in roots, stems and leaves of *A. mongholicus*. A total of four main isoflavonoids and four astragalosides (belong to triterpenoid saponins) were measured, and 44 differentially expressed genes (DEGs) of nine gene families, 44 DEGs of 16 gene families that encode for enzymes involved in isoflavonoid and triterpenoid saponin biosynthesis were identified, separately. Additionally, transcription factors (TFs) associated with isoflavonoid and triterpenoid saponin biosynthesis were analyzed, including 72 MYBs, 53 bHLHs, 64 AP2-EREBPs and 11 bZIPs. The above transcripts exhibit different expression trends in different organs.

**Conclusions:** Our study provides important genetic information for the essential genes of isoflavonoid and triterpenoid saponin biosynthesis in *A. mongholicus*, and provides a basis for developing its medicinal value.

**Background**

The perennial herbaceous genus *Astragalus mongholicus* Bunge, in the family Fabaceae (Leguminosae), contains many conventional therapeutic plants with a long history in traditional popular prescription China [1, 2]. *A. mongholicus* has been broadly utilized as a principal herbal in China, including Northeast, North, and Northwest of China, as well as in Mongolia and Korea [3]. Pharmacological studies and clinical practice have demonstrated that the dried roots of *A. mongholicus* have good potential therapeutic applications, including tonifying vital-energy, draining abscess, diuretic, reinforcing skin and tissue productive activities [4]. Recent studies reveal that its bioactivity is mainly assigned to isoflavonoids and triterpenoid saponins [5]. Isoflavonoids and triterpenoid saponins are widely distributed in plants. Hence, the biosynthesis pathway of them are well understood, especially in some model plants [6, 7]. However, *A. mongholicus* is a non-model organism with little available genomic information. And it seriously restricts the study of the regulatory mechanisms of development, biosynthesis of metabolites, and other physiological processes.

The biosynthesis pathway of isoflavonoids is a branch of the phenylpropanoid pathway [8, 9]. It is a class of phenolic secondary metabolites that can regulate the interactions of complex plant-microbial and are mainly synthesized in legumes [10, 11]. Phenylalanine ammonia-lyase (PAL) is the first key enzyme in the
phenylpropanoid pathway, cinnamate 4-hydroxylase (C4H) is the second enzyme, and 4-Coumarate CoA-ligase (4CL), which produce p-coumaroyl-CoA as a precursor for the synthesis of chalcone. Chalcone synthase (CHS) catalyzes p-coumaroyl-CoA to naringenin chalcone is the first committed step towards flavonoid biosynthesis. Then chalcone isomerase (CHI) catalyzes the cyclization of the chalcones, producing liquiritigenin or naringenin, which are the substrates for many downstream enzymatic reactions to form flavonoids in legume plants. The key step at the very beginning of the isoflavonoid metabolic pathway is isoflavone synthase (IFS) converts flavanones to their corresponding isoflavonoids.

Current research on the biosynthesis pathway of triterpenoid saponins can be divided into three stages, including the synthesis of upstream precursor substances, synthesis of the midstream carbocyclic skeleton, and formation of different types of triterpenoid saponins by various complex functionalization reactions downstream [12]. The upstream precursors isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) are mainly formed through the mevalonic acid pathway (MVA) and the methylerythritol pathway (MEP) [13–15]. The carbocyclic skeleton of triterpenoid saponins was synthesized by active DMAPP catalyzed by prenyltransferase and oxidosqualene cyclases (OSCs) [7, 16–17]. Finally, different types of triterpenoid saponin products were obtained by chemically modified cytochrome P450 monooxygenases (CYP450s), UDP-glycosyltransferases (UGTs) and glycosidases.

Transcriptome studies provide a useful perspective for the molecular biology of medicinal plants, including molecular mechanisms of gene function, cellular response, and different biological processes [18–22]. We combined next-generation sequencing (NGS) and single-molecule real-time (SMRT) sequencing to generate a more complete A. mongholicus transcriptome. And it has provided a valuable resource for studying the regulatory mechanisms of development, biosynthesis of metabolites, and other physiological processes of A. mongholicus without a reference genome. To further investigate the isoflavonoid and triterpenoid saponin biosynthesis of A. mongholicus, we determined the content of four primary isoflavonoids and four mainly astragalosides (belong to triterpenoid saponins) in roots, stems and leaves based on ultra-high performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS). Candidate key genes involved in isoflavonoid and triterpenoid saponin biosynthesis of A. mongholicus have been explored. Accordingly, this study provides novel insights into the isoflavonoid and triterpenoid saponin biosynthesis of A. mongholicus.

Results

**Accumulation of isoflavonoids and triterpenoid saponins in different organs of A. mongholicus**

Eight constituents that four came from isoflavonoid biosynthesis and four from triterpenoid saponins biosynthesis were quantitated in roots (AR), stems (AS) and leaves (AL) of A. mongholicus by comparison with standard compounds based on their retention times and MS fragmentation patterns. As shown in Fig. 1 four isoflavonoids and four triterpenoid saponins were mainly distributed in the roots.

**Transcriptome analysis of A. mongholicus combined sequencing approach**
To obtain as many high-quality unigenes as possible and differentiate the roots, stems, leaves transcriptome of *A. mongholicus*, a hybrid sequencing strategy that combined SMRT and NGS technology was taken. Firstly, nine mRNA samples from three different organs (roots, stems and leaves) were sequenced on the DNBSEQ platform. The raw data with low-quality sequences on ambiguous N base. After quality filtering, the total number of clean reads from the roots, stems and leaves in *A. mongholicus* were 132.46M, 133.18M, 129.68M, respectively (Table 1). Secondly, full-length transcripts were reconstructed by using the PacBio Sequel platform. In total, 14.21GB subreads were assembled, with a mean length of 1826 bp and an N50 length of 2191 bp (Additional file: Figure S1a). A total of 643,812 reads of insert (ROIs) were generated, and 305,998 were identified as full-length non-chimeric (FLNC) reads with a mean length of 1619 bp (Additional file: Figure S1b). After that, we applied Interactive Clustering and Error Correction (ICE) algorithm combined with the Quiver program for sequence clustering, and removed the redundant sequences via CD-Hit program, 121,107 non-redundant transcripts with a N50 value of 2,124 bp were yielded (Table 2; Additional file 1: Figure S1c.).

**Function annotation of full-length *A. mongholicus* transcriptome**

To obtain the putative functional annotation of *A. mongholicus* transcriptome, the set of 121,107 transcripts were annotated performed against a variety of protein databases (NR, NT, Swissprot, KEGG, KOG, Pfam and GO) (Table 3). There are 36,812 (30.40%) transcripts annotated by all databases in the seven databases and 104,756 (86.50%) transcripts annotated by any of the seven databases. The homologous species of *A. mongholicus* were analyzed by aligning sequences to the NR database (Fig. 2a) and showed that 30.31%, 16.13%, 8.67%, and 8.01% were mapped to the genes of *Cicer arietinum* (Leguminosae), *Medicago truncatula* (Leguminosae), *Trifolium subterraneum* (Leguminosae), *Glycine max* (Leguminosae).

There are 74,953 transcripts annotated into the GO terms, divided into three major categories: molecular function, cellular component and biological process (Fig. 2b). The highest proportion of the three GO categories mentioned above were “cellular process”, “cell”, and “catalytic activity”, respectively. In total, 77,196 transcripts were assigned to the KOG, which were classified into 25 functional clusters (Fig. 2c). Among them, “general function prediction” was the most highly, followed by “signal transduction mechanisms”, “posttranslational modification, protein turnover, chaperones”. Furthermore, in the KEGG annotation, 78,294 transcripts from *A. mongholicus* were mapped in the KEGG pathway, which involved five functional categories (Cellular Processes, Environmental Information Processing, Genetic Information Processing, Metabolism, Organismal Systems) (Fig. 2d). The “global and overview maps”, “carbohydrate metabolism”, and “translation” were the most representative pathways. Among them, 1156 transcripts were enriched in “phenylpropanoid biosynthesis”, 398 transcripts were annotated to “flavonoid biosynthesis”, and 170 transcripts were assigned to “isoflavonoid biosynthesis”. More importantly, 479 transcripts were annotated as related to “terpenoid backbone biosynthesis”, 173 transcripts were assigned to “sesquiterpenoid and triterpenoid biosynthesis” in *A. mongholicus* transcriptome.

**Identification of transcription factors**
Transcription factors (TFs) play a transient spatially regulated role in plant development and response to stress. For TFs prediction, we annotated 2455 TFs belonging to 53 different TF families from the transcriptome dataset of *A. mongholicus*. Among them, the C3H (219 gens), GRAS (189 gens), AP2-EREBP (178 gens), MYB (177 gens), bHLH (176 gens) families had the large members (Table 4). It was reported that TFs play vital roles in regulating isoflavonoid and terpene biosynthesis in many plant species, such as MYB, bHLH, AP2-EREBP, bZIP families [23-27]. By removing the TFs, which expression levels among three organs extremely lowly (FPKM>1), we identified 72 MYBs, 53 bHLHs, 64 AP2-EREBPs and 11 bZIPs. Among the 72 members in the MYB family, more than half of MYB genes were mainly expressed in roots. For 53 bHLH genes, 18 bHLHs had the highest expression level in roots, 16 bHLHs in stem and 19 bHLHs in leaves. For 64 members of the AP2-EREBP family, almost all AP2-EREBPs have a high expression level in roots. In addition, we observed that 5 of 11 bZIP genes were enriched in roots.

**Identification of organ-specific transcripts and differentially expressed genes**

A total of 58,896 transcripts were found to be simultaneously expressed in roots, stems and leaves (Additional file 2: Figure S2a). In addition, 6087 transcripts exhibited organ-specific expression, with 2124, 1897, and 2066 transcripts specifically expressed in roots, stems and leaves, respectively.

The differentially expressed genes (DEGs) among different organs (AL vs. AS, AR vs. AL, and AR vs. AS) were investigated, and the results were displayed in Fig. 3a. A comparison between the AR vs. AL revealed that 25,970 were DEGs, with 12,799 genes up-regulated, and 13,171 genes were down-regulated (Table 5). Moreover, roots and leaves had the most specifically expressed differential genes (4765) (Fig. 3a), suggesting a larger biological difference between roots and leaves. In AL vs. AS, 19,923 DEGs were identified, with 12,313 genes were up-regulated, and 7610 genes were down-regulated. A total of 18,951 DEGs were identified in the AR vs. AS comparison. Among them, 12,829 genes were up-regulated, and 6122 genes were down-regulated. In all comparison groups, 6035 genes were differentially expressed, and revealing that these genes play an important role in the metabolism among different organs of *A. mongholicus*.

KEGG (Fig. 3) and GO (Additional file 2: Figure S2b-d.) enrichment analysis with the DEGs among different organs were performed to analyze the identified transcripts further. In the different comparisons, the widest metabolism class involved “carbohydrate metabolism”, “amino acid metabolism”, “energy metabolism”, “lipid metabolism”, “biosynthesis of other secondary metabolites”.

**Identification of DEGs involved in isoflavonoid biosynthesis**

Isoflavonoids are identified almost exclusively in leguminous plants. The isoflavonoid biosynthetic in *A. mongholicus* was characterized by other legumes (Fig. 4) [28-31]. Total of 53 candidate genes with a FPKM value of more than one that encoded nine enzymes involved in isoflavonoid biosynthesis was identified, and an intense focus was set on 44 DEGs in this study discovered (Additional file 3: Table S1).
Isoflavonoids are synthesized from L-phenylalanine through the central phenylpropanoid pathway. The upstream of the isoavonoid synthesis pathway is the same as a flavonoid. This process of upstream metabolism involved 34 DEGs in *A. mongholicus* were identified, including 12 *AmPALs*, two *AmC4Hs*, three *Am4CLs*, five *AmCHSs*, and 12 *AmCHIs* (Fig. 4). Close to half of the *AmPALs* in roots and stems were higher than in leaves. *AmPAL6* exhibited the highest expression in roots among the three organs. *AmPAL10, AmPAL3, AmPAL11* showed higher expression in stems over roots and leaves. The expression level of *AmPAL4, AmPAL1* and *AmPAL7* in leaves were over the other two organs. *AmC4H1* showed the highest expression in stems, and *AmC4H2* was expressed predominantly in stems and rarely expressed in leaves. *Am4CL1* and *Am4CL2* had their highest expression in stems. *Am4CL3* had a high level in leaves, which differed from other *Am4CLs*. The expression level of *Am4CL1* in roots was higher than that of *Am4CL2* and *Am4CL3*. *AmCHS2* and *AmCHS3* were mainly expressed in roots. Notably, the expression level of *AmCHS2, AmCHS3* were higher than other *CHSs* in roots. *AmCHS4* and *AmCHS5* had their highest expression in stems, followed by leaves and roots. Among the 12 transcripts of *AmCHIs*, we found that more than half of them showed the highest expression level in roots. *AmCHI1, AmCHI12, AmCHI6, AmCHI9* were expressed predominantly in stems. In general, the phenylpropanoid biosynthesis pathway and the flavonoid biosynthesis pathway have a higher relative gene expression level in roots and leaves.

Isoflavonoid biosynthesis is a branch of flavonoid pathway. The enzymes involved in the isoavonoid pathway of *A. mongholicus* were as follows: one *AmIFSs*, five *AmHIDs*, one *Am4’-OMTs*, three *AmIF7GTs*. As shown in Fig. 4, all the candidate transcripts were investigated in detail and displayed differential expression levels in roots, stems and leaves of *A. mongholicus*. It is noteworthy that the vast majority of the candidate transcripts related to isoavonoid biosynthesis were more highly expressed in roots, over stems, and leaves, which were consistent with the contents of isoavonoids in *A. mongholicus*. The cooperative relationship between these candidate transcripts and the accumulation of isoavonoids implies that these genes have significant functions in isoavonoids biosynthesis of *A. mongholicus*.

### Analysis of DEGs in pathways related to triterpenoid saponins biosynthesis

A total of 58 candidate genes encoded 16 enzymes related to triterpenoid saponins biosynthesis were found by removing extremely lowly expressed genes with an FPKM value of less than 1. Among them, 44 candidate genes were considered as DEGs (Additional file 4: Table S2) and were the key candidates for discussion in our study.

The MVA pathway is the earliest and more traditional pathway for the synthetic biology of terpenoids. In the MVA pathway, 23 DEGs were identified (Fig. 5). The results showed that half of the candidate transcripts involved in MVA pathway displayed the highest expression levels in the roots of *A. mongholicus*. Based on our transcriptomic dataset, only one transcript for *AmHMGC*, *AmMVK*, and *AmPMK* were identified as DEGs involved in the MVA pathway. Moreover, they demonstrated the same regularity among the three organs and were showed the highest expression in stems, followed by roots or leaves. Of the six *AmAACTs*, *AmAACT6* and *AmAACT4* displayed higher expression levels in roots than that in the other two organs. However, the expression of *AmAACT3* and *AmAACT1* in roots was higher.
than other AmAACTs, especially AmAACT3. These suggest that AmAACT3 also plays an important role in the MVA pathway of A. mongholicus. HMGR is a key enzyme in the MVA pathway, 11 AmHMGRs were identified as DEGs of A. mongholicus. Notably, six AmHMGRs expression levels of roots were higher than that of stems and leaves. Compared with other AmHMGRs, AmHMGR5 was most highly expressed in roots, implied that the transcript played a vital role in the MVA pathway in roots of A. mongholicus. As far as three AmMVDs expression patterns were found in different organs, all of them were the highest in roots, followed by stems and leaves.

In our transcriptomic dataset, a total of 13 DEGs were found to be involved in the MEP pathway of A. mongholicus, including three AmDXSSs, two AmDXRs, two AmispDs, one AmispE, one AmgcpE, and four AmispHs. As shown in Fig. 5b, it was evident that almost all DEGs related to the MEP pathway were displayed the highest expression levels in leaves and least in roots. Both MEP and MVA pathways generate IPP and its isomer DMAPP, which are precursors of the production of terpenoids. Triterpenoids and sesquiterpenoids are synthesized by the MVA pathway, whereas monoterpenoids, diterpenoid, and tetraterpenoids are biosynthesized via the MEP pathway. In our study, the MVA pathway is mainly involved in the biosynthesis of triterpenoid saponins in A. mongholicus.

The biosynthesis of triterpenoid saponins from A. mongholicus mainly involves “terpenoid backbone biosynthesis” and “sesquiterpenoid and triterpenoid biosynthesis” pathways. Nine DEGs encoded for three enzymes were identified with regard to “sesquiterpenoid and triterpenoid biosynthesis” (Fig. 5d). AmFDPSs have two candidate transcripts, but only one gene was differentially expressed. AmFDPS showed similar expression among three organs, while the expression level in stems was highest. SQS and SQE are playing a pivotal role in the biosynthesis pathway of the carbon ring skeleton of triterpenoid saponins. Three AmSQEs exhibited higher expression in roots than in stems and leaves. The two AmSQSSs, AmSQE1 and AmSQE2, had the highest expression levels in stems, over roots and leaves.

In order to verify the gene expression level, five genes were randomly selected for a real-time qPCR (RT-qPCR) experiment. The results showed that the trend of transcript changes in these five gene transcriptomes was good with the quantitative results of qPCR. This result proves the accuracy of the transcriptome data (Additional file 5: Figure S3).

Discussion

Each technological change and advancement in sequencing technology has a profound impact on transcriptome research as a means to quickly, accurately, and efficiently interpret gene-level information from medicinal plants. Because of the high-throughput, sensitivity, accuracy, and low-cost advantages of NGS, it is widely used as a technical tool to reveal more information about the genes of medicinal plants [32–35]. Currently, transcriptome studies of A. mongholicus are mainly based on NGS [28, 30, 36]. The reference genome of A. mongholicus has not yet been published, resulting in a lack of genetic information about the plant. It is reported that SMRT gives a down to earth direct for the analysis of full-length transcripts in species lacking a genome assembly [37]. Moreover, SMRT has the advantages of
single-molecule real-time sequencing and super long reading length, which greatly makes up for the shortcomings of NGS [38, 39]. For all this, the problem of high error rate with SMRT sequencing can not be ignored. Therefore, the combination of NGS and SMRT has been widely recognized for the analysis of genes involved in the biosynthesis of secondary metabolites in plants [40–42]. We combined NGS and SMRT sequencing to generate higher quality and more complete/full-length of *A. mongholicus* transcriptome. In our study, a total of 643,812 CCS reads were generated, yielding 121,107 non-redundant transcript isoforms with an N50 value of 2124 bp. Based on these highly accurate transcripts, 104,756 (86.50%) transcripts were successfully annotated by any of the seven databases. Additionally, 2455 TFs (belonging to 53 different TF families) were also obtained. We yielded 6035 common differentially expressed transcripts in all comparison groups by analyzing gene expression levels in different tissues.

Isoflavonoids, such as formononetin, ononin, calycosin, calycosin-7-glucoside, are important bioactive compounds in *A. mongholicus*. Based on UPLC-MS/MS analysis, the contents of four isoflavonoids in roots, stems and leaves were compared, and the results showed that accumulations of four isoflavonoids were various in different tissues. However, there was a common trend that the four isoflavonoids especially accumulate in roots. This tendency to accumulate active secondary compounds in *A. mongholicus* is consistent with previous reports [43–44]. To reveal the relationship between the isoflavonoid biosynthesis and accumulation in different plant organs of *A. mongholicus*, candidate genes responsible for the isoflavonoid biosynthesis must be identified and characterized in the biosynthetic pathway. KEGG pathway annotation revealed that 1156 transcripts were enriched in “phenylpropanoid biosynthesis”, 398 transcripts were annotated to “flavonoid biosynthesis”, and 170 transcripts were assigned to “isoflavonoid biosynthesis”. In the present study, we found that the same gene family transcripts may exhibit different expression trends. Some key genes that the expression patterns were consistent with the accumulation patterns of isoflavonoids in different organs, including three *PALs* (*AmPAL2, AmPAL6, AmPAL12*), two *CHSs* (*AmCHS2, AmCHS3*), seven *CHIs* (*AmCHI2, AmCHI3, AmCHI4, AmCHI5, AmCHI7, AmCHI10, AmCHI11*), one *IFS* (*AmIFS*), three *HIDs* (*AmHID1, AmHID2, AmHID3*), one 4’-*OMT* (*Am4’-OMT1*), and three *IF7GTs* (*AmIF7GT1, AmIF7GT2, AmIF7GT3*). It implied that these genes have significant functions in isoflavonoids biosynthesis of *A. mongholicus*. In addition, it is noteworthy that more genes in the upstream of isoflavonoid synthesis pathway were expressed at high levels in stems or leaves. In contrast, most of the genes in the middle and downstream pathways were expressed at high levels in the roots. We further analyze that the aboveground part of *A. mongholicus* might be the important site for the biosynthesis of the phenylpropanoid pathway and flavonoids biosynthetic pathway, while the underground part is the important biosynthesis site of isoflavonoids.

Triterpenoid saponins are important plant secondary metabolites with various biological activities in plants and widely distributed in dicotyledonous plants [45]. In our transcriptome data, 479 transcripts were annotated as related to “terpenoid backbone biosynthesis”, 173 transcripts were assigned to “sesquiterpenoid and triterpenoid biosynthesis”. The distribution and accumulation of four triterpenoid saponins (astragaloside I, astragaloside II, astragaloside III and astragaloside IV) in the studied different organs of *A. mongholicus* were showed in Fig. 1. The determination results showed that the contents of triterpenoid saponins in the roots were highest. The number of differentially expressed transcripts
encoded for 16 enzymes involved in triterpenoid saponins biosynthesis of *A. mongholicus* was 44. Some genes related to triterpenoid saponins biosynthesis were more highly expressed in roots than stem and leaves. Notably, *HMGCRs* is the key enzyme in the MVA pathway [46], and *AmHMGCR1, AmHMGCR2, AmHMGCR3, AmHMGCR5, AmHMGCR6, AmHMGCR11* expressed highest in roots, which was in accordance with the pattern of triterpenoid saponins levels in different organs. Triterpenoid saponins are mainly synthesized by the MVA pathway [47]. Beyond that, the *SQS* families and *SQE* families possess many members in plants, and some of them are key enzymes that regulate plant saponin synthesis [48]. *AmSQE1, AmSQE2, AmSQE3* exhibited the highest expression in root among three organs. Hence, the above transcripts may have important roles in triterpenoid saponins biosynthesis of *A. mongholicus* and could potentially be used as target markers in *A. mongholicus* breeding programs aimed at increasing astragalosides production.

It has been reported that MYB, bHLH, AP2-EREBP, bZIP families can regulate isoavonoid and triterpenoid saponin biosynthesis. We identified 72 MYB unigenes, 53 bHLH unigenes, 64 AP2-EREBP unigenes and 11 bZIP unigenes. Interestingly, more than half of MYBs and almost all AP2-EREBPs were mainly expressed in root, which showed the highest accumulation of isoflavones and triterpenoid saponins in *A. mongholicus*. It is suggested that MYBs and AP2-EREBPs may play an important role in regulating the organ-specific biosynthesis of isoflavonoids and triterpenoid saponins of *A. mongholicus*. During the regulation of plant secondary metabolic pathways, the regulation of TFs mostly belongs to positive regulation, but the secondary metabolic network is complex with more branches of metabolic. In order to increase the metabolic flow of synthetic target compounds more efficiently, it is effective to inhibit by-product branch metabolism by using technical methods [49–50]. Therefore, attention should be paid to MYBs and AP2-EREBPs in *A. mongholicus*, which can more effectively regulate the synthesis of isoflavonoids and astragalosides.

**Conclusions**

Isoflavonoids and triterpenoid saponins at roots, stems, leaves of *A. mongholicus* were tested and carried out the analysis of *A. mongholicus* transcriptome with the combination of short-read NGS and long-read SMRT sequencing in this study. Our data list more accurate and complete transcripts in two important biosynthesis pathways of *A. mongholicus* than previous studies. We also identified putative key transcripts and transcription factors involved in isoflavonoid and triterpenoid saponin biosynthesis. Comprehensive analysis of transcriptomics and metabolomics revealed the gene expression and metabolite profiles across organs, suggesting organ-specific biosynthesis, accumulation and modification of isoflavonoid and triterpenoid saponin may occur in different tissues of *A. mongholicus*. Therefore, the work provided provides valuable basics for future research on gene discovery, molecular breeding, and metabolic engineering in *A. mongholicus*.

**Materials And Methods**

**Plant materials**
Fresh roots (AR), stems (AS), leaves (AL) of *A. mongholicus* were collected from Anhui University of Chinese Medicine, Anhui Province, China. Three independent biological replicates for each sample were washed clean, surface dried, and flash-frozen in liquid nitrogen, then stored at -80°C until extraction of chemical constituents and RNA.

**RNA extraction, quantification and sequencing**

Total RNA was extracted from fresh samples with an improved CTAB method according to the official operating manuals. The RNA quality was evaluated using a NanoDrop and an Agilent 2100 Bioanalyzer (Thermo Fisher Scientific, MA, USA). A DNBSEQ platform conducted NGS sequencing. Oligo (dT)-attached magnetic beads were used to purified mRNA. Purified mRNA was fragmented into small pieces with a fragment buffer at the appropriate temperature. Then first-strand cDNA was generated using random hexamer-primed reverse transcription, followed by second-strand cDNA synthesis. Afterward, A-Tailing Mix and RNA Index Adapters were added by incubating to end repair. PCR amplified the cDNA fragments obtained from the previous step. The double-stranded PCR products of the previous step were heated denatured and circularized by the splint oligo sequence to get the final library. Then, raw reads were further filtered to obtain clean reads by removing reads containing adaptors, poly-N or low quality. For SMRT sequencing, the first-strand cDNA was synthesized from 800-1000 ng total RNA by SMARTer PCR cDNA Synthesis Kit (Clontech, Japan). Second-strand cDNA synthesis was performed using SMARTScribe™ Reverse Transcriptase (Clontech, Japan). A sufficient amount of double-stranded cDNA was produced by large-scale PCR with PrimeSTAR GXL DNA Polymerase (Clontech, Japan). The double-stranded cDNA was size-selected into 0-5 kb. The PacBio Sequel sequencer performed sequencing reactions. Library preparation and sequencing were conducted by BGI (Shenzhen, Guangdong province, China).

**Analysis of the Iso-Seq data**

Raw sequencing data produced by Pacific Bioscience Sequel were processed through the SMRT analysis package. Raw Polymerase reads that had full passes >0 and the predicted consensus accuracy >0.75 were selected for producing Reads of Insert (ROI). ROIs with a minimum length of 300 bp were classified into full-length non-chimeric and non-full-length transcript sequences based on whether the 5′ primer, 3′ primer, and poly-A tail were observed or not. The full-length sequences were processed to consensus isoforms by Iterative Clustering for Error Correction (ICE) algorithm and then polished via Quiver quality-aware algorithm. To improve the accuracy of full-length transcripts, the additional nucleotide errors were further corrected using DNBSEQ RNA-Seq reads by Long-Read De Bruijn Graph Error Correction (LoRDEC) [51]. The consensus isoforms with high quality (the expected Quiver accuracy $\geq 0.95$) from each library were merged and redundancy was removed using CD-HIT [52] (parameter: -c 0.98 -T 6 -G 0 -aL 0.90 -AL 100 -aS 0.98 -AS 30) based on the sequence similarity to obtain final unique full-length isoforms.

Final full-length isoforms were mapped to the NR (non-redundant protein sequences), NT (non-redundant nucleotide sequence), SwissProt (a manually annotated and reviewed protein sequence database), KEGG (Kyoto Encyclopedia of Genes and Genomes, version 59), KOG (Clusters of Eukaryotic Orthologous
Groups) and Pfam (Protein families) by Blast software (version 2.2.23) [53] with default parameters (under a threshold E-value ≤10−5) to get the isoform annotations. GO (Gene Ontology) annotations and functional classifications were obtained using Blast2GO program (version 2.5.0, E-value ≤10−5) [54] based on NR annotations. Transcription factors (TF) were predicted by mapping to PlntfDB.

**Differentially expression genes (DEGs) analysis**

All clean reads from all samples of *A. mongholicus* were mapped to full-length transcriptome using Bowtie2 (v2.2.5) [55], then the expression level of unigenes was calculated by RSEM (v1.2.12) [56]. FPKM (fragments per kb per million fragments) values were used to quantify each unigene. Differentially expressed genes (DEGs) were identified using DEseq2 [57] with Q value (adjust P-value) < 0.001 and fold change ≥ 2 or ≤ -2. The identified DEGs were enriched to GO term and KEGG database with Phyper in the R package by Q value <= 0.05. The FPKM values of DEGs were greater than or equal to 1.

**Quantitative real-time PCR (RT-qPCR) analysis**

Five randomly selected genes from the transcriptome database were validated by RT-qPCR. The primer sequences and actin gene in these experiments were listed in (Additional file 6: Table S3). Total RNA from roots, stems, and leaves of *A. mongholicus* were reversely transcribed into the first-strand cDNA using PrimeScript II 1st Strand cDNA Synthesis Kit (Takara, Japan). RT-qPCR for genes were carried out with three biological replicates and three technical replicates by TB Green® Premix Ex Taq™ II (Takara, Japan) on Agilent StrataGene Mx3000P qPCR System (Agilent, USA). The 2^{-\Delta\Delta CT} method was used to analyze the relative expression levels of genes.

**Determination of isoflavonoids and triterpenoid saponins in *A. mongholicus***

The dried and homogeneous powder (0.2g) of roots, stems and leaves isolated from *A. mongholicus* was accurately weighted with 4 ml 75%. Then, ultrasonication was performed at room temperature for 60 min, and the supernatant was then collected. The solution was filtered through a 0.22 μm membrane, and the clear supernatant was injected into the UPLC-MS/MS system for analysis.

Chromatographic separation was performed on an ExionLC AD (SCIEX, USA), and an Acquity UPLC BEH C18 column (100 mm×2.1 mm, 1.7 mm) and a C18 Pre-column (2.1 mm × 100 mm, 1.7 μm; Waters) was applied for all analyses. The gradient elution utilizing acetonitrile as solvent A and water (0.1% formic acid) as solvent B with a flow rate of 0.2 mL/min was performed. The gradient was as follows: 0-2 min (8%40%A), 2-5 min (40%-55%B), 5-7 min (55%-60%B), 7-8 min (60%-80%B), 8-10 min (80%-8%B). The column temperature was maintained at 28℃, and the equilibration of 2 min was applied between individual runs.

An AB SCIEX Triple Quad 4500 (AB SCIEX, USA) equipped with an electrospray ionization source (ESI) was used for the detection of the MS signals. The MS detection conditions were used in positive ion multiple reaction monitoring (MRM) modes with the following parameters: the capillary voltage operated at 5.5 kV, the source temperature at 550 °C, nebulizer gas: 55 psi; curtain gas: 35 psi.
Abbreviations

CCS
Circular consensus sequencing; NR:NCBI non-redundant protein; NT:Nucleotide sequences; GO:Gene Ontology; KEGG:Kyoto Encyclopedia of Genes and Genomes; KOG:Eukaryotic Ortholog Groups; Pfam:Protein family; PAL:Phenylalanine ammonialyase; C4H:Trans-cinnamate 4-monooxygenase ; 4CL:4-coumarate CoA ligase; CHS:Chalcone synthase; CHI:Chalcone isomerase ; IFS:2-hydroxyisoflavanone synthase; HID:2-hydroxyisoflavanone dehydratase; 4'-OMT:2,7,4'-trihydroxyisoflavanone; IF7GT:Isoflavone 7-O-glucosyltransferase; I3'H:Isoflavone 3'-hydroxylase; I2'H:Isoflavone 2'-hydroxylase; AACT:Acetyl-CoA C-acetyltransferase ; HMGCS:Hydroxymethylglutaryl-CoA synthase; HMGCR:Hydroxymethylglutaryl-CoA reductase; MVK:Mevalonate kinase; PMK:Phosphomevalonate kinase; MVD:Diphosphomevalonate decarboxylase; DXS:1-deoxy-D-xylulose-5-phosphate synthase; DXR:1-deoxy-D-xylulose-5-phosphate reductoisomerase ; ispD:2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase ; ispE:4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; ispF:2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase ; gcpE:(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase; ispH:4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase; IDI:Isopentenyl-diphosphate Delta-isomerase ; GPS:Geranylgeranyl diphosphate synthase ; FDPS:Farnesyl diphosphate synthase; SQS:Squalene synthase; SQE:Squalene epoxidase; CAS:Cycloartenol synthase; UGT:Glucuronosyltransferase

Declarations

Authors’ contributions

MZY, LPZ and HSP conceived and designed the experiments. MZY, LPZ, SSC, TYS performed the experiments. MYZ wrote the manuscript and analyzed the data and prepared the figures and tables. All the authors read and approved the manuscript.

Author details

1 College of Pharmacy, Anhui University of Chinese Medicine, Hefei 230012, China

2 Research Unit of DAO-DI Herbs, Chinese Academy of Medical Sciences, Beijing 100700, China

3 Institute of Conservation and Development of Traditional Chinese Medicine Resources, Anhui Academy of Chinese Medicine, Hefei 230012, China

4 Anhui Province Key Laboratory of Research & Development of Chinese Medicine, Hefei 230012, China

Acknowledgments

We would like to thank the BGI Institute for assistance with experiments.

Funding
This research was supported by the National Natural Science Foundation of China (81573543, 81773853, 82073957-81703633), and CAMS Innovation Fund for Medical Sciences (2019-I2M-5-065).

Availability of data and materials

The sequencing data of all sample-sequencing results have been submitted to the NCBI SRA (PRJNA681346) (https://dataview.ncbi.nlm.nih.gov/object/PRJNA681346?reviewer=cf9fmfe1blgqlaej16ksehqjrt).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References


37. Li J, Lee YH, Denton MD, Feng QJ, Rathjen JR, Qu ZP, Adelson DL. Long read reference genome-free reconstruction of a full-length transcriptome from Astragalus membranaceus reveals transcript


### Tables

**Table 1** Summary of transcriptome sequencing data from roots, stems and leaves in *A. mongholicus*

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw Reads (M)</th>
<th>Clean Reads (M)</th>
<th>Clean Bases (Gb)</th>
<th>Clean Reads Q20 (%)</th>
<th>Clean Reads Q30 (%)</th>
<th>Clean Reads Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR-1</td>
<td>43.92</td>
<td>43.81</td>
<td>6.57</td>
<td>96.04</td>
<td>87.34</td>
<td>99.75</td>
</tr>
<tr>
<td>AR-2</td>
<td>44.99</td>
<td>44.85</td>
<td>6.73</td>
<td>96.04</td>
<td>87.48</td>
<td>99.7</td>
</tr>
<tr>
<td>AR-3</td>
<td>43.92</td>
<td>43.80</td>
<td>6.57</td>
<td>96.06</td>
<td>87.47</td>
<td>99.73</td>
</tr>
<tr>
<td>AS-1</td>
<td>43.92</td>
<td>43.81</td>
<td>6.57</td>
<td>96.15</td>
<td>87.63</td>
<td>99.75</td>
</tr>
<tr>
<td>AS-2</td>
<td>43.92</td>
<td>43.80</td>
<td>6.57</td>
<td>95.91</td>
<td>87.02</td>
<td>99.73</td>
</tr>
<tr>
<td>AS-3</td>
<td>45.67</td>
<td>45.57</td>
<td>6.84</td>
<td>96.84</td>
<td>88.83</td>
<td>99.78</td>
</tr>
<tr>
<td>AL-1</td>
<td>42.16</td>
<td>42.07</td>
<td>6.31</td>
<td>96.19</td>
<td>87.72</td>
<td>99.77</td>
</tr>
<tr>
<td>AL-2</td>
<td>43.92</td>
<td>43.82</td>
<td>6.57</td>
<td>96.10</td>
<td>87.50</td>
<td>99.77</td>
</tr>
<tr>
<td>AL-3</td>
<td>43.92</td>
<td>43.79</td>
<td>6.57</td>
<td>95.91</td>
<td>87.11</td>
<td>99.71</td>
</tr>
</tbody>
</table>

Note: AR Roots, AS Stems, AL Leaves
Table 2 The non-redundant transcripts statistics

<table>
<thead>
<tr>
<th>Total number</th>
<th>Total length</th>
<th>N50</th>
<th>N90</th>
<th>Max length</th>
<th>Min length</th>
<th>Sequence GC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>121,107</td>
<td>212,470,804</td>
<td>2124</td>
<td>1087</td>
<td>17,059</td>
<td>197</td>
<td>41.09%</td>
</tr>
</tbody>
</table>

Table 3 Annotation of transcripts against seven public databases

<table>
<thead>
<tr>
<th>Annotated databases</th>
<th>Total</th>
<th>NR</th>
<th>NT</th>
<th>Swissprot</th>
<th>KEGG</th>
<th>KOG</th>
<th>Pfam</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcript Number</td>
<td>121,107</td>
<td>98,388</td>
<td>95,310</td>
<td>75,812</td>
<td>78,294</td>
<td>77,196</td>
<td>59,554</td>
<td>74,953</td>
</tr>
<tr>
<td>Percentage</td>
<td>100%</td>
<td>81.24%</td>
<td>78.70%</td>
<td>62.60%</td>
<td>64.65%</td>
<td>63.74%</td>
<td>49.17%</td>
<td>61.89%</td>
</tr>
</tbody>
</table>

Table 4 The number of 20 transcription factors (TFs) with different expression in different organs
<table>
<thead>
<tr>
<th>TF family</th>
<th>Number of transcripts</th>
<th>Number of up-regulated transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AR vs. AL</td>
</tr>
<tr>
<td>C3H</td>
<td>219</td>
<td>21</td>
</tr>
<tr>
<td>GRAS</td>
<td>189</td>
<td>19</td>
</tr>
<tr>
<td>AP2-EREBP</td>
<td>178</td>
<td>17</td>
</tr>
<tr>
<td>MYB</td>
<td>177</td>
<td>22</td>
</tr>
<tr>
<td>bHLH</td>
<td>176</td>
<td>48</td>
</tr>
<tr>
<td>WRKY</td>
<td>117</td>
<td>31</td>
</tr>
<tr>
<td>G2-like</td>
<td>90</td>
<td>30</td>
</tr>
<tr>
<td>NAC</td>
<td>85</td>
<td>12</td>
</tr>
<tr>
<td>ARF</td>
<td>73</td>
<td>4</td>
</tr>
<tr>
<td>ABI3VP1</td>
<td>69</td>
<td>9</td>
</tr>
<tr>
<td>C2H2</td>
<td>68</td>
<td>11</td>
</tr>
<tr>
<td>mTERF</td>
<td>68</td>
<td>18</td>
</tr>
<tr>
<td>Trihelix</td>
<td>65</td>
<td>23</td>
</tr>
<tr>
<td>C2C2-Dof</td>
<td>53</td>
<td>6</td>
</tr>
<tr>
<td>FAR1</td>
<td>52</td>
<td>6</td>
</tr>
<tr>
<td>MADS</td>
<td>52</td>
<td>20</td>
</tr>
<tr>
<td>SBP</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td>TUB</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>FHA</td>
<td>37</td>
<td>14</td>
</tr>
<tr>
<td>C2C2-CO-like</td>
<td>36</td>
<td>30</td>
</tr>
</tbody>
</table>

**Table 5** Number of up-regulated and down-regulated transcripts between trial groups: AR vs. AL, AR vs. AS and AL vs. AS
Additional Files

Additional File 1: Figure S1. Sequencing results of Pacbio sequel platform. a Length distribution of subreads. b The number and length distributions of FLNC reads. c The number and length distributions of non-redundant transcript isoforms.

Additional File 2: Figure S2. Statistics of transcripts and differentially expressed genes. a Venn diagram of transcripts in different organs. b, c, d Scatterplot of the GO functional enrichment of DEGs (AR vs. AL, AR vs. AS and AL vs. AS, respectively)

Additional File 3: Table S1. Differentially expressed genes involved in isoflavonoid biosynthesis.

Additional File 4: Table S2. Differentially expressed genes involved in triterpenoid saponins biosynthesis.

Additional File 5: Table S3. Primers used in validation experiment of gene expression by RT-qPCR.

Additional file 6: Figure S3. RT-qPCR validation of 5 randomly selected genes in different organs. The actin gene was Am18S.